International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 4, May-June 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

House Price Estimates Based on Machine Learning Algorithm

Jakir Khan[1], Dr. Ganesh D[2]

[1]Student, [2]Assistant Professor,
[1,2]Department of MCA, School of CS & IT, Jain University, Bangalore, Karnataka, India

ABSTRACT

Housing prices are increasing every year, necessitating the creation of a long-term housing price strategy. Predicting a home's price will assist a developer in determining a home's purchase price, as well as a consumer in determining the best time to buy a home. The sale price of real estate in major cities depends on the specific circumstances. Housing prices are constantly changing from day to day and are sometimes fired rather than based on estimates. Predicting real estate prices by real factors is a key element as part of our analysis. We want to make our test dependent on all of the simple metrics that are taken into account when deciding the significance. In this research we use linear regression techniques pathway and our results are not self-inflicted process rather is a weighted method of various techniques to give the most accurate results. There are fifteen features in the data collection. In this research. There has been an effort to build a forecasting model for determining the price based on the variables that influence the price.The results have proven to be effective lower error and higher accuracy than individual algorithms are used.

KEYWORDS: *Machine Learning, Linear Regression algorithm*

How to cite this paper: Jakir Khan | Dr. Ganesh D "House Price Estimates Based on Machine Learning Algorithm" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.795-799, URL: www.ijtsrd.com/papers/ijtsrd42367.pdf

IJTSRD42367

I. INTRODUCTION

Identification House prices continue to change for the day and going out for a day and sometimes smoking rather than based on estimates. Machine learning algorithms are used for modeling, in which the machine learns data and uses what it has learned to predict new data. The most well-known paradigm of forecasting research is backstage. The proposed model for accurately predicting future outcomes has economic, financial, banking, health, commercial, recreational, sport and other sectors. Many variables are used in a single way to predict house prices. Forecasting houses prices with real features are our main crux research project. We use various how to get back on track, too our results are not self-determined process rather is a weighty method of various techniques to give the most accurate results. The results have proven to be effective lower error and higher accuracy than individual algorithms are used. The aim of this research is to gain a useful understanding of the housing market in the United States by analyzing the actual historical data of the transaction. Looking for models that can estimate the worth of a house based on a set of features. Practical models can help home buyers and for real estate brokers to make informed decisions. In addition, it can assist in predicting home rates in the future and the process of formulating housing market policy. In comparison to other traditional methods, our work can achieve a better performance by experimenting with real-world property transaction data. With our research project, consider USA as our main and predictable location prices of real estate in various locations around USA.

II. LITERATURE REVIEW

A large number of scholars have participated to this work on the analysis of house price prediction in the past. The following is a review of the text related to the study of house price prediction.

In paper **[1]** 4,000 unprocessed datasets from eight counties were collected from real estate agents' Services with Multiple Listings (MLS) in Washington, DC. Three separate learning algorithms was put to the test the hedonic hypothesis (PCR, SVR and K-NN). For component analysis and decomposition, PCA was used. The Chi-Square Quantile-Quantile plot and Henze-Multivariate Zirkler's Normality Test were used to perform the normality test. Model performance has shown that PCR has side effects for SVR and K-NN. The suitability and replacement of PCR, SVR, and K-NN in the application of the hedonic pricing policy was also confirmed in this report.

In paper **[2]** investigating price redistribution a combination of Sale Price for each variety and present many new variables For example, Fig. 1 shows the log conversion Individual price distribution neighbors. The purpose of the engineering features is to improve data familiarity and equity, while setting the highest parameter Iteration times are used to improve data consistency.

In paper **[3]** to improve interpretation and improve performance of predictive models, data reduction strategies like Stepwise and Boosting are exploited to get more important predictions. In addition, PCA, data conversion technique, used to find things that are important to combine with SVM. There are 34,857 observations and 21 variables in the original dataset. From 2016 to 2018, each observation depicts a real sold house transaction in Melbourne. The aim of this research is to analyze real history a transaction database to gain valuable insights into real estate at a market

in downtown Melbourne. Looking for useful models to predict that a house's worth is determined by its location given a set of its features.

In paper **[4]** estimate each parameter according to its value in determining the pricing of the system and this has led us to increase the value saved by each parameter system. We rated 3 different machines algorithms to learn and test our system with a different combination that can guarantee the best possibly the reliability of our results. A program that aims to provide accurate predictions housing prices have been built. The program makes good use of Linear Regression, Forest retreat, magnified magnitude. Efficiency of algorithm added more with the use of Neural Networks. consider it Mumbai as our main and predictable location prices of real estate in various locations around Mumbai consider a certified database .

In paper **[5]** develop the industrial revolution's fourth stage over time with the growing popularity of in big data technology, it had moved importance and rapid development in the field of data science. The dataset contain 171,155 property transaction record from 2013 to 2017. Because there are missing data of features. The proposed in-convolution learning model provides a better solution to real estate price forecasts with higher accuracy and faster integration rate, and that the proposed prediction process can increase the robustness of the convolutional neural network model.

In paper **[6]** There are five standard machine learning processes in place used in this study namely Random Forest Regressor, Decision

Tree Director, Ridge, Lasso and Regular Linear. Before confirming the predictable results of each, appropriate algorithm suspensions are identified first based on training database by to call the best estimator method. The dataset contain 15 columns of 4066 records, which 3252 for training and 814 for validation.

In paper **[7]** from June 1996 to August 2014, the data preparation process resulted in 39,554 housing transactions. This collection program can be done inside a single element or by combining features in a very high-end aircraft. Noble (2006) suggests that the details are more easily separated from high-resolution spaces because there is a kernel function that can be applied to any data set divided equally.

However, locating the kernel is a difficult challenge function.

In paper **[8]** study provides insight into machine learning skills as well geographical methods of molding in complex urban areas using a huge volume of information from many geospatial sources. There are21,928 homes, 125,000 house portraits, and approximately 470,000 street view pictures.

In paper **[9]** with a simple drop in line we try to minimize error, and in SVR we try to match his error within something the limit. It is a retrieval algorithm and uses the same Support Vector Machines (SVM) retrieval method analysis. The train data set consists of 11200 records with 9 explanatory variables. In test data set, there were around 1480 records with 9 variables.

In paper **[10]** more than 300,000 data points are included in this dataset, which includes 26 variables that reflect housing prices exchanged between 2009 and 2018.Analysis of experimental data is an important step before building a retrospective model. In this way, the researcher scans to find data patterns, which helps to select appropriate methods for machine learning. Three types of Machine Learning methods include Random Forest, XGBoost, and LightGBM and two machine learning techniques including. Three types of Machine Learning methods include Random Forest, XGBoost, and LightGBM and two machine learning techniques including proven to be effective in a variety of computer-aided applications (He et al., 2016)

III. EXISTING MODEL

Our findings show how multiple geo-data sources can be used in machine learning applications to demonstrate housing pricing patterns and policy makers to assist in understanding human settlements. The sale price of real estate in major cities depends on the specific circumstances. Although a macro viewpoint of the real estate appreciation rate could be more useful, consistency of performance is not the only metric to consider when determining the right model.

IV. PROPOSED MODEL

The proposed model for accurately predicting future outcomes has applications in real state. The aim of this statistical study is to aid in our understanding of the the house-to-house partnership features and how these variables are used to forecast house price. To develop the proposed method for calculating rates of increase in house prices generalisation and replicability. The aim of this research is to through analyzing a real historical transactional dataset to derive valuable insight into the housing market in USA. Our dataset contains a wide range of critical parameters, and data mining is at the heart of our system. We started by cleaning up the entire dataset and truncating the outlier values. Similarly, we weighted each parameter based entirely on its importance in determining the system's pricing, which resulted in an increase in the weight that each parameter retains in the system. We choose three outstanding system learning algorithms and tested our system with excellent combinations that will ensure the accuracy and consistency of our effects [5].Even after that, we used a totally new approach to improve precision. Even after that, we took a novel approach to improve the quality of our survey, which revealed that the real estate fee is often influenced by local facilities such as a train station, grocery store, college, pharmacy, temple, parks, and so on. And now we'd like to present our specific approach to addressing this need. We use the Google Maps API to narrow down a distance of 0.5 km based on locality seek. Now, if we discover certain public places in the circle, the value of the belongings rises in proportion. We tested this with manual scenarios, and the results were excellent in terms of prediction accuracy [8].

V. SYSTEM ARCHITECTURE

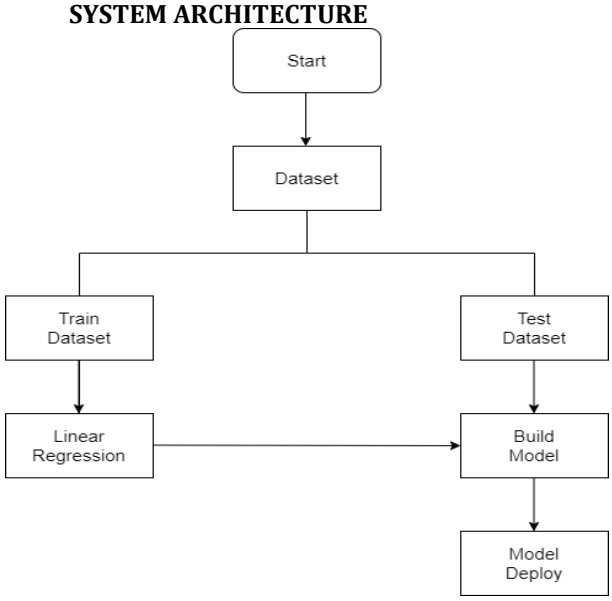


Figure:V.I

The structure of the structure in this way, which includes several important elements as an algorithm. Line reversal is a basic method of prediction. A linear regression algorithm will be used to estimate house price based on available data obtained from image use.

VI. METHODOLOGY

In this research paper method is available in the Python Scikit Library-Learn to do a grid search efficient use of hyper-parameter tuning in a given machine learning algorithm. This is a useful way for an inexperienced data scientist to get suggestions for configuring parameters in selected algorithms.

A. Dataset Exploration

The data set is split into two sections, The separation has been aligned to 75:25 ratio of training and testing. Includes converting raw data to data loader which is then used for training of the model and getting insight into our dataset. In this process, various steps involved like converting the data to data frame, visualizing the dataset, splitting of data frame into training and validation.

A.1 Features Description Table

Name	Type	Description
Id	Numerical	Uniq Id
Date	Numerical	Sold Time
Price	Numerical	House Price (prediction outcome)
Bedroom	Numerical	No. of bedroom
Bathroom	Numerical	No. of bathroom
Sqft_living	Numerical	Living size
Sqft_lot	Numerical	Lot size
Floors	Numerical	No. of floor
Waterfront	Numerical	Waterfront size
Condition	Categorical	Type of house
grade	Numerical	Rating of house
Sqft_above	Numerical	Size
Sqft_basement	Numerical	Size
Yr_built	Numerical	Built Year(2015-16)

Table: A.1.1

A.1 Training vs Validation Accuracy Curve

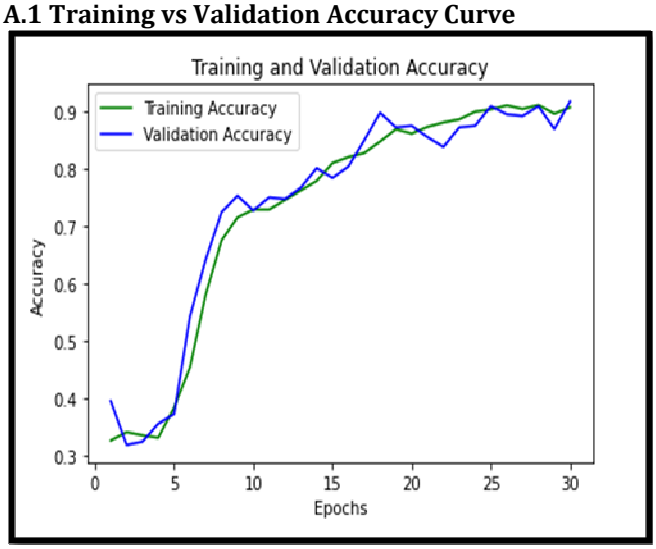


Figure: A.1.1

A.2 Training vs Validation Loss Curve

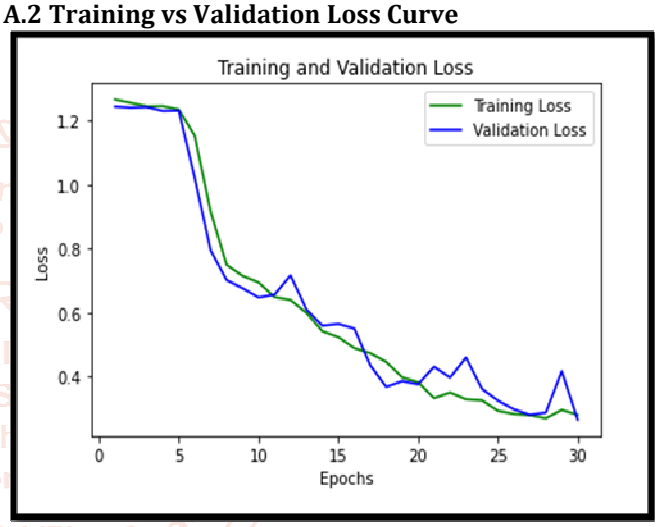


Figure: A.2.1

A.3 Dataset collection

The dataset is collected from https://www.kaggle.com/harlfoxem/housesalesprediction. Between May 2014 and May 2015, homes were sold in the area. A good collection of data to test the basic concept of simple regression models. Before preparing the data we need to explore the data first. The dataset has 16204 number which is use to train the model. It has number 5404 which is used for testing.

A.4Data cleansing and exploratory analysis: Check is there any null value.

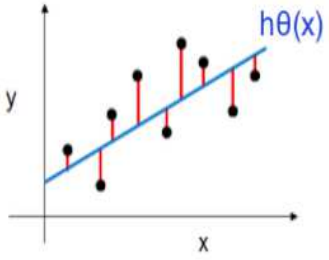
#check for nulls in the data
houses.isnull().sum()

```
Out[3]: id               0
        date             0
        price            0
        bedrooms         0
        bathrooms        0
        sqft_living      0
        sqft_lot         0
        floors           0
        waterfront       0
        view             0
        condition        0
        grade            0
        sqft_above       0
        sqft_basement    0
        yr_built         0
        yr_renovated     0
        zipcode          0
        lat              0
        long             0
        sqft_living15    0
        sqft_lot15       0
        dtype: int64
```

Figure: A.4.1

B. Model defining

Creating a Linear Regression model, training function and optimizing the hyper parameters like learning rate, number of epochs. simple linear regression is a mathematical method of modeling the correlation between the predictor X and the Y response variable. It assumes that there may be a direct correlation between these two variables and we use that to expect a limited output. simple line deceleration is a completely simple way to do supervised mastering. but, while it may be the end of the road and the most straightforward, it is an important starting point for all the regression strategies, so it is important to fully understand what it is about for miles. It is also widely used and smooth translation: it helps to gain a better understanding of the connection between feedback and prediction. Mathematically, we are able to document these specific relationships as: y = -0 + β1x + e; therey output flexibility (also called response, targeting or systematic flexibility). e.g. housing costsx input variables (also called active, descriptive or non-descriptive) e.g. the height of the house is square meters β0 capture (price y when x = 0)β1 is the coefficient x and the slope of the return line ("normal boom in Y associated with the increase of one unit in X")e is a time of error when using linear regression, the algorithm produces the best live path using the coefficients version zero and β1, such as miles as close as possible to the actual facts (reduces the ratio of distances twice between all data and road). as soon as we find ero zero and β1 we will use the model to wait for a response.



- The blue line — line of first-class stable — is the line that minimises the number of squared defects.
- The black circles are the found values of x and y (actual facts).
- The errors (or residuals) are the vertical distances between the discovered values (the real statistics) and the pleasant suit lane.
- the blue line's slope is 1; the intercept (the value of y when x=0) is 0The most straightforward method of estimation is linear regression. It uses two variables as variables: a predictor variable and a variable that is the most important one first, if the predictor variable and su. This regression estimates are used to explain the relationship between one known variable and one or more unobservable variables. The components describe the equation of the regression equation with one dependent and one unbiased variable [6].b = y + x*a, where a = ranking on the independent variable, y = normal, x = regression coefficient, and b = approximate existing variable score.

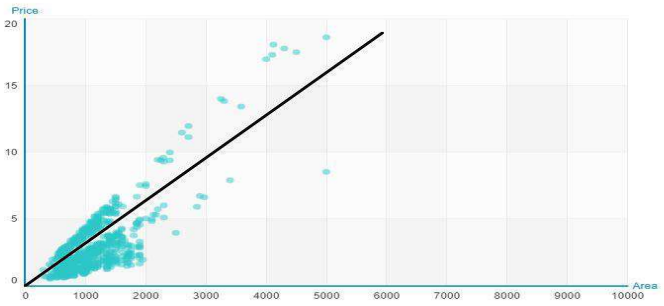


Figure: B.1 Linear regression scatter plot

C. Implementation

The implementation is done in Python 3. To develop and train the Linear Regression model, the Ker as library, which operates on top of Tensor flow, is used. The Anaconda package manager is used to import all dependencies.

D. Classification

Includes prediction on the test dataset, visualizing the performance of our model and seeing how our model is performing.

D.1 Any correlations between variables:

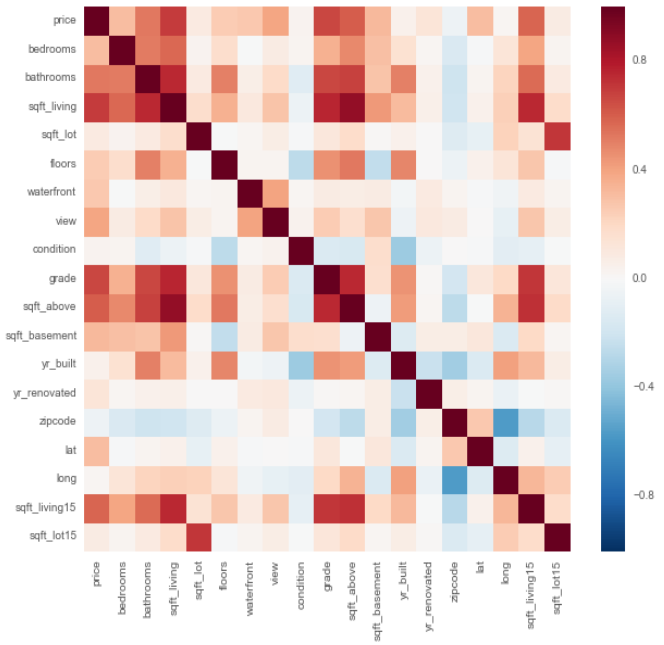


Figure:D.1.1

VII. RESULT

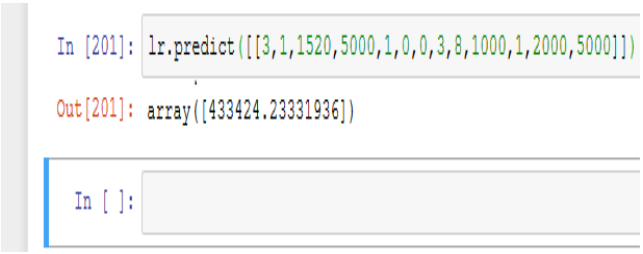


Figure: VII.I Final outcome

VIII. CONCLUSION

A system has been developed that seeks to provide reliable forecasts for housing prices, and allows for the optimal use of the Linear Regression algorithm. The use of neural networks has improved algorithm performance even more. Customers will be content because the device have the right results and eliminate the chance of finding the wrong structure. In asset analysis, new mechanical calculation methods can be used. Compared to a simple model, we can use one to predict house price.

IX. FUTURE WORK

The machine's accuracy should be improved. As the gadget's scale and computing power grow, it would be possible to add even more cites. Furthermore, we will incorporate exceptional UI/UX techniques for improved simulation of the outcomes in a more interactive way by the use of Augmented truth [eleven]. Also, a mastering device may be developed to collect user feedback and documents so that the gadget would display the most suitable effects to the user based on his choices.

Acknowledgment

I should to pass on my genuine propensity and commitment to Dr MN. Nachappa and Asst. Prof: Dr Ganesh D and undertaking facilitators for their compelling steerage and steady motivations all through my evaluation work. Their optimal bearing, total co-activity and second insight have made my work productive.

References

[1] Oladunni, T. and Sharma, S., 2016. Hedonic Housing Theory — A Machine Learning Investigation. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA),*.

[2] Lu, S., Li, Z., Qin, Z., Yang, X. and Goh, R., 2017. A hybrid regression technique for house prices prediction. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),*.

[3] Phan, T., 2018. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE),*.

[4] Varma, A., Sarma, A., Doshi, S. and Nair, R., 2018. House Price Prediction Using Machine Learning and Neural Networks. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT),*.

[5] Piao, Y., Chen, A. and Shang, Z., 2019. Housing Price Prediction Based on CNN. *2019 9th International Conference on Information Science and Technology (ICIST),*.

[6] [6]Masrom, S., Mohd, T., Jamil, N., Rahman, A. and Baharun, N., 2019. Automated Machine Learning based on Genetic Programming: a case study on a real house pricing dataset. *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS),*.

[7] Ho, W., Tang, B. and Wong, S., 2020. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), pp.48-70.

[8] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F. and Ratti, C., 2020. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, p.104919.

[9] Manasa, J., Gupta, R. and Narahari, N., 2020. Machine Learning based Predicting House Prices using Regression Techniques. *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),*.

[10] Truong, Q., Nguyen, M., Dang, H. and Mei, B., 2020. Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174, pp.433-442